Introductory Statistics

Dr. Jarad Niemi

STAT 4610X - Iowa State University

January 30, 2025

Outline

- Binomial
- Poisson
- Normal

Binomial

A binomial random variable Y is the count of the number of successes out of n attempts where each attempt is

- independent and
- has a probability of success π.

We write

$$Y \sim Bin(n, \pi).$$

You should recall the following properties of a binomial distribution

- $Im[Y] = \{0, 1, 2, \ldots, n\}$,
- $E[Y] = n\pi$, and
- $Var[Y] = n\pi(1 - \pi)$.

Binomial example

Season Totals

NAME	MIN	FGM	FGA	FTM	FTA	3PM	3PA	PTS	OR	DR	REB	AST	TO	STL	BLK
Curtis Jones G	618	127	281	54	66	54	142	362	8	84	92	48	25	29	4
Keshon Gilbert G	643	110	213	74	101	17	52	311	15	59	74	92	57	29	1
Joshua Jefferson F	567	91	172	66	84	10	34	258	35	127	162	56	35	37	12
Tamin Lipsey G	591	73	151	41	55	20	62	207	20	29	49	59	30	45	7
Milan Momcilovic F	372	55	118	14	18	31	70	155	9	44	53	12	9	5	5
Dishon Jackson C	372	59	102	66	86	0	0	184	40	60	100	10	21	14	21
Nate Heise G	382	30	69	7	11	8	33	75	13	41	54	20	13	22	3
Brandton Chatfield F	290	26	46	18	27	0	1	70	39	30	69	6	12	8	11
Nojus Indrusaitis G	76	10	28	8	14	2	13	30	0	4	4	3	4	2	0
Demarion Watson G	88	8	15	6	9	2	6	24	8	15	23	3	3	3	6
Kayden Fish F	15	1	4	1	2	0	0	3	1	2	3	0	1	0	0
JT Rock C	15	1	1	0	0	0	0	2	1	3	4	0	0	0	1
Cade Kelderman G	20	1	4	0	0	0	2	2	1	0	1	3	1	3	0
Conrad Hawley F	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total		592	1204	355	473	144	415	1683	222	527	749	312	216	197	71

Binomial inference

When collecting binomial data, we are interested in making statements about the probability of success π. The most useful statement is an uncertainty interval for π. In introductory statistics courses, we teach a confidence interval based on the Central Limit Theorem:

$$\hat{\pi} = y/n, \quad \hat{\pi} \pm z_{a/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}.$$

where $z_{a/2}$ is the z-critical value such that the interval has (frequentist) probability of a to contain the true value π. In this course, we will just use $2 \approx 1.96$ so that the interval has approximately 95% (frequentist) probability. But, nobody actually uses this formula.

Binomial uncertainty intervals

```
y    <- 54
n    <- 66
phat <- y/n
phat + c(-1, 1) * 2 * sqrt(phat * (1 - phat) / n) # Introductory statistics

[1] 0.7232304 0.9131333

binom.test(y, n)$conf.int                          # Exact confidence interval

[1] 0.7039345 0.9023648
attr(,"conf.level")
[1] 0.95

prop.test( y, n)$conf.int                          # Better approximate interval

[1] 0.7000550 0.8985865
attr(,"conf.level")
[1] 0.95

qbeta(c(.025, .975), 1 + y, 1 + n - y)             # Bayesian credible interval

[1] 0.7080366 0.8924405
```

Poisson distribution

A Poisson random variable Y is the count of the number of successes where there is no clear upper maximum. The count is typically over some time, space, or space-time. We write

$$Y_i \stackrel{ind}{\sim} Po(\lambda).$$

where λ is the rate of occurrence and ind indicates that each observation (i) is independent. Please remember the following properties of a Poisson distribution

- $Im[Y_i] = \{0, 1, 2, \ldots\}$,
- $E[Y_i] = \lambda$, and
- $Var[Y_i] = \lambda$.

Game Log

2024-25

2024-25 Season (ISU)

DATE	OPP	RESULT	MIN	FG	FG%	3PT	3P%	FT	FT%	REB	AST	BLK	STL	PF	TO	PTS
Mon 1/27	@ ARIZ	L 86-75 OT	41	1-11	9.1	0-8	0.0	6-8	75.0	5	2	0	2	4	3	8
Sat 1/25	@ ASU	W 76-61	40	10-22	45.5	5-10	50.0	8-10	80.0	7	1	0	3	0	2	33
Tue 1/21	vs UCF	W 108-83	34	8-19	42.1	2-7	28.6	1-2	50.0	8	2	0	1	1	1	19
Sat 1/18	@ WVU	L 64-57	36	8-16	50.0	1-6	16.7	1-1	100.0	8	0	0	0	3	4	18
Wed 1/15	vs 9 KU	W 74-57	36	9-17	52.9	5-6	83.3	2-2	100.0	6	1	1	2	1	4	25
Sat 1/11	@ TTU	W 85-84 OT	35	8-15	53.3	3-7	42.9	7-8	87.5	1	0	0	2	1	0	26
Tue 1/7	vs UTAH	W 82-59	32	10-17	58.8	3-7	42.9	0-0	0.0	5	6	0	2	2	0	23
Sat 1/4	vs 25 BAY	W 74-55	28	6-13	46.2	2-5	40.0	0-0	0.0	3	2	0	4	1	2	14
Mon 12/30	@ COLO	W 79-69	29	5-16	31.3	2-7	28.6	8-8	100.0	3	0	0	1	2	1	20
Sun 12/22	vs MORG	W 99-72	30	7-15	46.7	3-9	33.3	2-2	100.0	2	6	2	2	1	0	19
Sun 12/15	vs OMA	W 83-51	24	2-9	22.2	0-5	0.0	0-0	0.0	2	6	0	3	0	1	4
Thu 12/12	@ IOWA	W 89-80	32	8-15	53.3	5-8	62.5	2-2	100.0	6	0	0	0	1	2	23
IOWA CORN CY-HAWK SERIES																
Sun 12/8	vs JKST	W 100-58	25	6-12	50.0	5-10	50.0	2-4	50.0	1	5	0	1	2	1	19
Wed 12/4	vs 5 MARQ	W 81-70	28	6-14	42.9	2-7	28.6	0-1	0.0	5	1	0	1	1	1	14
BIG 12-BIG EAST BATTLE																
Wed 11/27	vs COLO	W 99-71	26	7-14	50.0	3-9	33.3	2-2	100.0	6	3	0	1	1	0	19
THE MAUI INVITATIONAL PRESENTED BY NOVAVAX - 5TH PLACE GAME																
Tue 11/26	vs DAY	W 89-84	31	6-13	46.2	3-6	50.0	4-4	100.0	2	1	0	0	3	0	19
THE MAUI INVITATIONAL PRESENTED BY NOVAVAX																
Mon 11/25	vs 4 AUB	L 83-81	29	4-11	36.4	2-6	33.3	4-4	100.0	5	3	0	1	2	0	14
THE MAUI INVITATIONAL PRESENTED BY NOVAVAX																
Mon 11/18	vs IUIN	W 87-52	27	7-14	50.0	4-10	40.0	2-3	66.7	6	1	0	0	1	2	20
Mon 11/11	vs KC	W 82-56	29	7-11	63.6	3-5	60.0	3-3	100.0	5	5	0	1	0	0	20
Mon 11/4	vs MVSU	W 83-44	26	2-7	28.6	1-4	25.0	0-2	0.0	6	3	1	2	1	1	5

Poisson inference

When collecting Poisson data, we are interested in making statements about the rate λ. The most useful statement is an uncertainty interval for λ. A Central Limit Theorem based interval is

$$\hat{\lambda} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\lambda} \pm z_{a/2}\sqrt{\hat{\lambda}/n}.$$

But, nobody actually uses this formula.

Poisson uncertainty intervals

```
y         <- c(5,7,8,8,6,1,5,3,3,2,2,6,1,5,6,2,5,6,5,6)
lambdahat <- mean(y)
n         <- length(y)

lambdahat + c(-1, 1) * 2 * sqrt(lambdahat / n)      # CLT interval

[1] 3.640834 5.559166

exp(confint(glm(y ~ 1, family = "poisson")))        # Poisson regression style

   2.5 %   97.5 %
3.722916 5.605016

qgamma(c(.025, .975), 1 + sum(y), 1 + n)            # Bayesian credible interval

[1] 3.574429 5.372849
```

The interpretation is average rebounds per game.

Poisson process

A Poisson process is a random variable Y is the count of the number of successes over some amount of time, space, or space-time (T). We write

$$Y \stackrel{ind}{\sim} Po(\lambda T).$$

where λ is the rate of occurrence. Please remember the following properties of a Poisson distribution

- $Im[Y] = \{0, 1, 2, \ldots\}$,
- $E[Y] = \lambda T$, and
- $Var[Y] = \lambda T$.

Poisson process example

Season Totals

NAME	MIN	FGM	FGA	FTM	FTA	3PM	3PA	PTS	OR	DR	REB	AST	TO	STL	BLK
Curtis Jones G	618	127	281	54	66	54	142	362	8	84	92	48	25	29	4
Keshon Gilbert G	643	110	213	74	101	17	52	311	15	59	74	92	57	29	1
Joshua Jefferson F	567	91	172	66	84	10	34	258	35	127	162	56	35	37	12
Tamin Lipsey G	591	73	151	41	55	20	62	207	20	29	49	59	30	45	7
Milan Momcilovic F	372	55	118	14	18	31	70	155	9	44	53	12	9	5	5
Dishon Jackson C	372	59	102	66	86	0	0	184	40	60	100	10	21	14	21
Nate Heise G	382	30	69	7	11	8	33	75	13	41	54	20	13	22	3
Brandton Chatfield F	290	26	46	18	27	0	1	70	39	30	69	6	12	8	11
Nojus Indrusaitis G	76	10	28	8	14	2	13	30	0	4	4	3	4	2	0
Demarion Watson G	88	8	15	6	9	2	6	24	8	15	23	3	3	3	6
Kayden Fish F	15	1	4	1	2	0	0	3	1	2	3	0	1	0	0
JT Rock C	15	1	1	0	0	0	0	2	1	3	4	0	0	0	1
Cade Kelderman G	20	1	4	0	0	0	2	2	1	0	1	3	1	3	0
Conrad Hawley F	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total		**592**	**1204**	**355**	**473**	**144**	**415**	**1683**	**222**	**527**	**749**	**312**	**216**	**197**	**71**

Poisson process inference

When collecting Poisson process data, we are interested in making statements about the rate λ. The most useful statement is an uncertainty interval for λ. A Central Limit Theorem based interval is

$$y/T \pm z_{a/2}\sqrt{(y/T)/T}.$$

But, nobody actually uses this formula.

Poisson process uncertainty intervals

```
y <- 92                                  # Number of rebounds
t <- 618                                 # Number of minutes played

y/t + c(-1, 1) * 2 * sqrt(y/t / t)       # CLT interval

[1] 0.1178263 0.1799083

qgamma(c(.025, .975), 1 + y, 1 + t)      # Bayesian credible interval

[1] 0.1212649 0.1822776
```

These intervals are interpreted per minute played.

Poisson process per game

```
r <- c(5,7,8,8,6,1,5,3,3,2,2,6,1,5,6,2,5,6,5,6) # Rebounds per game
y <- sum(y)                                       # Total rebounds
t <- length(r)                                    # Total games played

y/t + c(-1, 1) * 2 * sqrt(y/t / t)                # CLT interval

[1] 3.640834 5.559166

qgamma(c(.025, .975), 1 + y, 1 + t)               # Bayesian credible interval

[1] 3.574429 5.372849
```

These intervals are interpreted per game played.

Normal

A normal random variable Y is a continuous random variable We write

$$Y_i \stackrel{ind}{\sim} N(\mu, \sigma^2).$$

with mean μ and variance σ^2 (or standard deviation σ). You should recall the following properties of a normal distribution

- $Im[Y_i] = (-\infty, \infty) = \mathbb{R}$,
- $E[Y_i] = \mu$, and
- $Var[Y] = \sigma^2$.

Normal inference

When collecting normal data, we are (typically) interested in making statements about the mean μ. The most useful statement is an uncertainty interval for μ. In introductory statistics courses, we teach the confidence interval

$$\overline{y} \pm z_{a/2} \times s/\sqrt{n}$$

where

- $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean and
- $s = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ is the sample variance.

We do actually use this formula!!

Normal uncertainty intervals

```
# Fictitious data that matches Aldritch Potgeiter's driving distance
y    <- rnorm(12, mean = 0, sd = 20)
y    <- y - mean(y) + 328.7
ybar <- mean(y)
n    <- length(y)
s    <- sd(y)

ybar + c(-1, 1) * 2 * s / sqrt(n) # Approximation using 2

[1] 318.0891 339.3109

t.test(y)$conf.int                # Exact and Bayesian interval

[1] 317.0228 340.3772
attr(,"conf.level")
[1] 0.95
```

This is the uncertainty around his mean driving distance.

Summary

The building blocks of many statistical analyses are the following probability distributions:

- Binomial (count with a known upper maximum)
- Poisson (count with no known upper maximum)
- Normal (not a count)

In this slide set, we introduced some basic uncertainty intervals for using data to make statements about parameters in these models.